Supplementary material

A Dataset documentation and intended uses

The evaluation suite is intended to be used for a more fine-grained evaluation of Natural Language Generation models. All the source datasets have an associated data card with detailed explanations on what they contain and how they can be read: `https://gem-benchmark.com/data_cards`.

Our 80 challenge sets are detailed in the *Changes to the Original Dataset for GEM* section of the data card of each concerned dataset.

A collaborative repository that will allow anyone to submit new challenge sets for NLP test suites in general has also been set up, where the code developed to create our evaluation suite can also be found: `https://github.com/GEM-benchmark/NL-Augmenter`.

B URL to platform where the benchmark can be viewed and downloaded

All our datasets are (and will remain) downloadable from Hugging Face: `https://huggingface.co/datasets/viewer/?dataset=gem`. The datasets of our presented evaluation suite all have the *challenge* prefix in the *Split* unfolding menu.

C Licensing

All datasets and code can be downloaded freely for research purposes; licensing details specific to each dataset can be found in the *Licensing Information* section of each data card (see Section A).

The authors bear all responsibility in case of violation of rights regarding the created datasets.

D Additional information on the ToTTo Named Entity splits

The categories within gender, ethnicity, and nationality were chosen based on data availability; The ToTTo dataset includes mostly tables that do not focus on people. As a result, only seven people in the original test set are marked as having a non-binary gender. Similar sparsity informed the grouping of nationalities by continent – only 19 countries are represented by more than 10 people in the test set. In case a person has citizenships across multiple continents, we may include the person in any of the included continents.

Finally, ethnicity is very sparsely annotated in WikiData; only 150 test examples in ToTTo have this information and 128 of these are African Americans. We thus are unable to compare the performance on, e.g., Yoruba or Punjabi people, both of which have fewer than five instances. Another caveat here is that only 21 of the 128 people are female. We thus compare the African American population to results on a subset that includes all US citizens.

E Additional information on the syntactic complexity scale

Sentences in the WikiAuto test sets were annotated with one of the following developmental levels: (0) simple sentences, including questions (1) infinitive or -ing complement with subject control;

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

	mT5 base		mT5 small		mT5 large		mT5 xl	
Subset	Vocab	MSTTR	Vocab	MSTTR	Vocab	MSTTR	Vocab	MSTTR
XSUM	4682	0.72	3954	0.67	4531	0.71	4629	0.73
WebNLG (en)	1862	0.62	2020	0.58	1726	0.63	1807	0.65
WebNLG (ru)	2655	0.71	2284	0.68	2759	0.73	2625	0.72
Wikilingua (ru)	7843	0.43	6062	0.31	10410	0.50	12815	0.60
Wikilingua (es)	12859	0.41	9298	0.30	13988	0.48	19494	0.59
Wikilingua (tr)	3399	0.58	3365	0.57	3555	0.58	3362	0.58
Wikilingua (vi)	5100	0.45	3406	0.28	6796	0.56	7531	0.60
Czech restaurants	465	0.53	492	0.54	566	0.56	666	0.59
E2E	133	0.28	137	0.28	131	0.28	236	0.28
SG-dialog	4169	0.68	4052	0.67	4326	0.69	4186	0.68
MLSUM (de)	35717	0.78	35351	0.77	35224	0.78	37096	0.78
MLSUM (es)	31073	0.71	29271	0.70	28969	0.71	30567	0.71

Table 1: Vocabulary size, and the mean-segmental type-token (MSTTR) ratio for each of our models, for all of the main datasets.

(2) conjoined noun phrases in subject position; conjunctions of sentences, of verbal, adjectival, or adverbial construction; (3) relative or appositional clause modifying the object of the main verb; nominalization in object position; finite clause as object of main verb; subject extraposition; (4) subordinate clauses; comparatives; (5) nonfinite clauses in adjunct positions; (6) relative or appositional clause modifying subject of main verb; embedded clause serving as subject of main verb; nominalization serving as subject of main verb; (7) more than one level of embedding in a single sentence.

F Diversity of Outputs

Table 1 presents the vocabulary size, and the mean-segmental type-token (MSTTR) ratio for our models, based on the test sets of our main datasets. Both measures reflect the diversity of the output. We observe that, although there are some exceptions, the mT5-small model tends to have a smaller vocabulary than mT5-large, which in turn tends to have a smaller vocabulary than mT5-xl. The same is true for the MSTTR values. This means that larger models tend to have more variation in their output, which may make these models less repetitive in the eyes of their users.

G More Results by complexity

Tables 2 and 3 and Figures 1–3 show additional results on the effect of different complexity aspects of inputs, expanding on those presented in Section 6. In all three cases, data-to-text, simplification, and dialog modeling, the complexity measures do not show any meaningful (negative) correlation with various performance metrics. While there can be multiple other confounders and proxy measures for complexity, it is interesting to observe that models do not always struggle with examples that humans would find more challenging. Instead, other measures like train-test overlap, often have a much stronger and consistent influence on the model performance.

Figure 4 is an extended version of Figure 3 of the main paper and additionally shows the behaviours of the models on seen and unseen data splits, and on the subpopulations based on the occurrence of a given entity in both a subject and an object positions of properties of the same input. At first sight, English outputs seem more affected than the Russian ones (c-f), but this is largely due to the fact that there are extremely few unseen predicates and entities in the Russian data (e.g. (c) *unseen* for Russian is based on one single unseen property). In Russian, there are also extremely few cases (3) of an entity being both subject and object of a property, so (f) is not very meaningful in this language. All the details about instance numbers in each subpopulation can be found on the respective data cards (`https://gem-benchmark.com/data_cards`).

	mT5 base			mT5 small			mT5 large			mT5 xl		
Subset	BLEURT	BLEU	Length	BLEURT	BLEU	Length	BLEURT	BLEU	Length	BLEURT	BLEU	Length
Val	0.46	66.24	21.56	0.46	65.44	22.21	0.46	65.89	21.81	0.45	65.20	21.76
Test	0.03	42.17	25.20	-0.15	35.14	28.21	0.16	46.46	24.94	0.22	49.07	24.27
1 pred	0.20	47.34	10.47	-0.04	38.56	11.18	0.29	52.44	10.29	0.37	57.83	10.13
2 preds	0.08	42.72	18.11	-0.10	35.39	20.12	0.21	46.99	17.85	0.30	53.84	17.26
3 preds	0.03	40.05	25.00	-0.18	30.70	29.81	0.13	44.19	24.07	0.20	47.41	23.43
4 preds	-0.04	40.18	31.86	-0.23	32.21	36.24	0.12	44.52	30.81	0.17	47.20	29.75
5 preds	-0.10	41.18	36.22	-0.26	35.48	40.38	0.08	45.24	36.80	0.10	45.26	35.48
6 preds	-0.10	41.75	42.23	-0.19	39.88	44.07	0.06	45.92	42.61	0.09	46.05	41.53
7 preds	-0.14	43.24	46.20	-0.23	42.91	49.67	-0.00	48.96	48.48	0.00	48.36	48.80

Table 2: BLEURT and BLEU scores, along with the prediction lengths for the different baseline models, applied to different subsets of the English part of the WebNLG dataset.

	mT5 base			mT5 small			mT5 large			mT5 xl		
Subset	BS	BLEU	Length	BS	BLEU	Length	BS	BLEU	Length	BS	BLEU	Length
Val	0.90	17.14	10.52	0.89	17.94	10.23	0.90	17.53	10.06	0.90	17.00	9.46
Test	0.90	19.25	10.39	0.89	15.74	10.93	0.90	16.82	12.02	0.89	17.37	10.98
1 pred	0.85	5.34	8.14	0.84	3.55	10.66	0.85	2.76	14.79	0.84	5.32	11.41
2 preds	0.91	21.26	9.76	0.90	14.57	9.60	0.91	22.31	9.40	0.91	26.23	8.54
3 preds	0.90	20.76	11.24	0.90	20.20	11.34	0.91	20.91	11.41	0.90	18.01	10.96
4 preds	0.92	18.49	13.71	0.92	16.72	13.60	0.92	20.93	15.76	0.92	19.19	17.26
5 preds	0.92	23.65	14.67	0.92	23.06	17.11	0.92	26.41	17.33	0.91	15.26	15.56

Table 3: Bertscore (BS) and BLEU scores, along with the prediction lengths for the different baseline models, applied to different subsets of the Czech restaurant dataset.

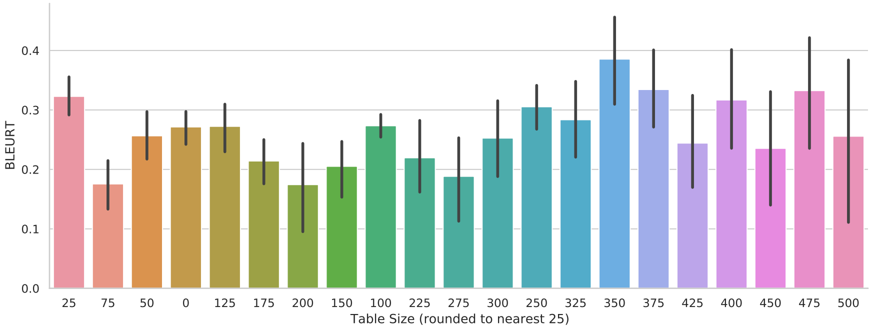

Figure 1: The effect of the input table size on model performance in ToTTo is relatively minor.

H Extended Transformation results

In Figure 5, we expand the findings presented in Figure 2 of the main paper to all transformation sets. We can see that the findings are consistent across all six types of transformations, varying only slightly in terms of their diversity results. In all cases, the performance-related metrics (except for BLEURT) drop significantly, indicating that models struggle with these examples. In the case of spelling mistakes (bfp02, bfp05) and scramble, there is an increase in vocabulary size, indicating a stronger reliance on the language model part of the encoder-decoder model. However, in many cases the local recall also increased, indicating a higher fraction of words that are copied verbatim from the input.

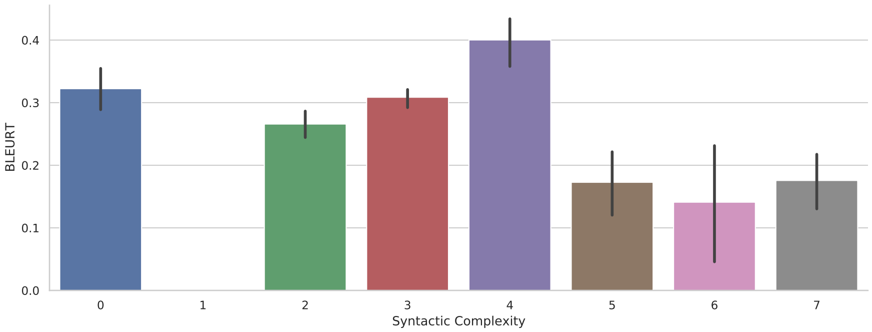

Figure 2: The effect of the input complexity on BLEURT Score in a simplification task is only apparent for very complex inputs (clases 5-7). Note that no example with complexity class 1 was found in the TURK/ASSET test sets.

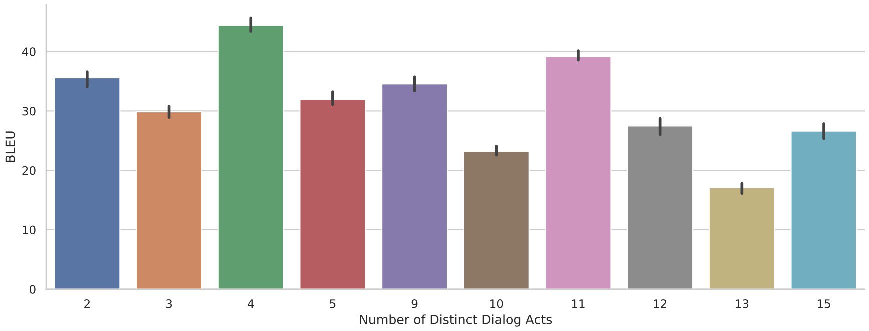

Figure 3: The effect of the number of dialog acts that need to be verbalized on the BLEU Score in the schema-guided dialog dataset. There is no consistent decrease as we would have expected.

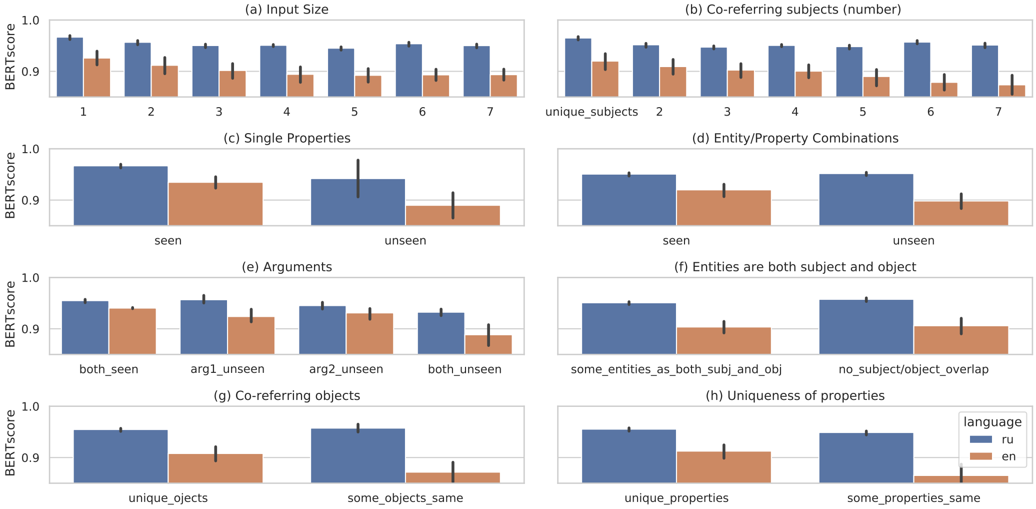

Figure 4: WebNLG results for English and Russian for all subpopulations. The scores of the four models are averaged; error bars indicate variance between model sizes.

Original Input Source	**Backtranslated Input Source**
In its pure form , dextromethorphan lives as a white powder.	Dextromethorph in its pure form lives as a white powder.
Alessandro Mazzola is an Italian former football player.	Alessandro Mazzola is a former Italian football player.
He died six weeks later on January 13th, 888.	He died six weeks later on 13 January, 888.

Table 4: Paired examples of the original input source and a corresponding backtranslation sample generated from the backtranslation model we use. All examples above are from the TURK test set of the Wiki-Auto Asset benchmark for text simplification.

I Analyzing Backtranslation Outputs

In Table 4, we can observe some of the changes backtranslation performs. In the second and third examples, we can see that the paraphrases change the order of adjectives appearing before the respective nouns. However, backtranslation can also perform larger, non-local changes - such as in the first example where it converts a topicalized source sentence to a non-topicalized one. Backtranslation can also perform changes simultaneously at many points in the sentence, such as in the example (not in the table) *It is* ***situated*** *at the coast of the Baltic Sea ,* ***where it encloses the city of Stralsund****.* → *It is* ***located*** *on Baltic Sea coast ,* ***where the historic town of Stralsund is founded****.* In this example, we can see it performing several diverse changes such as synonyms (situated → located) and phrases (where it encloses the city of Stralsund → where the historic town of Straslund is founded).

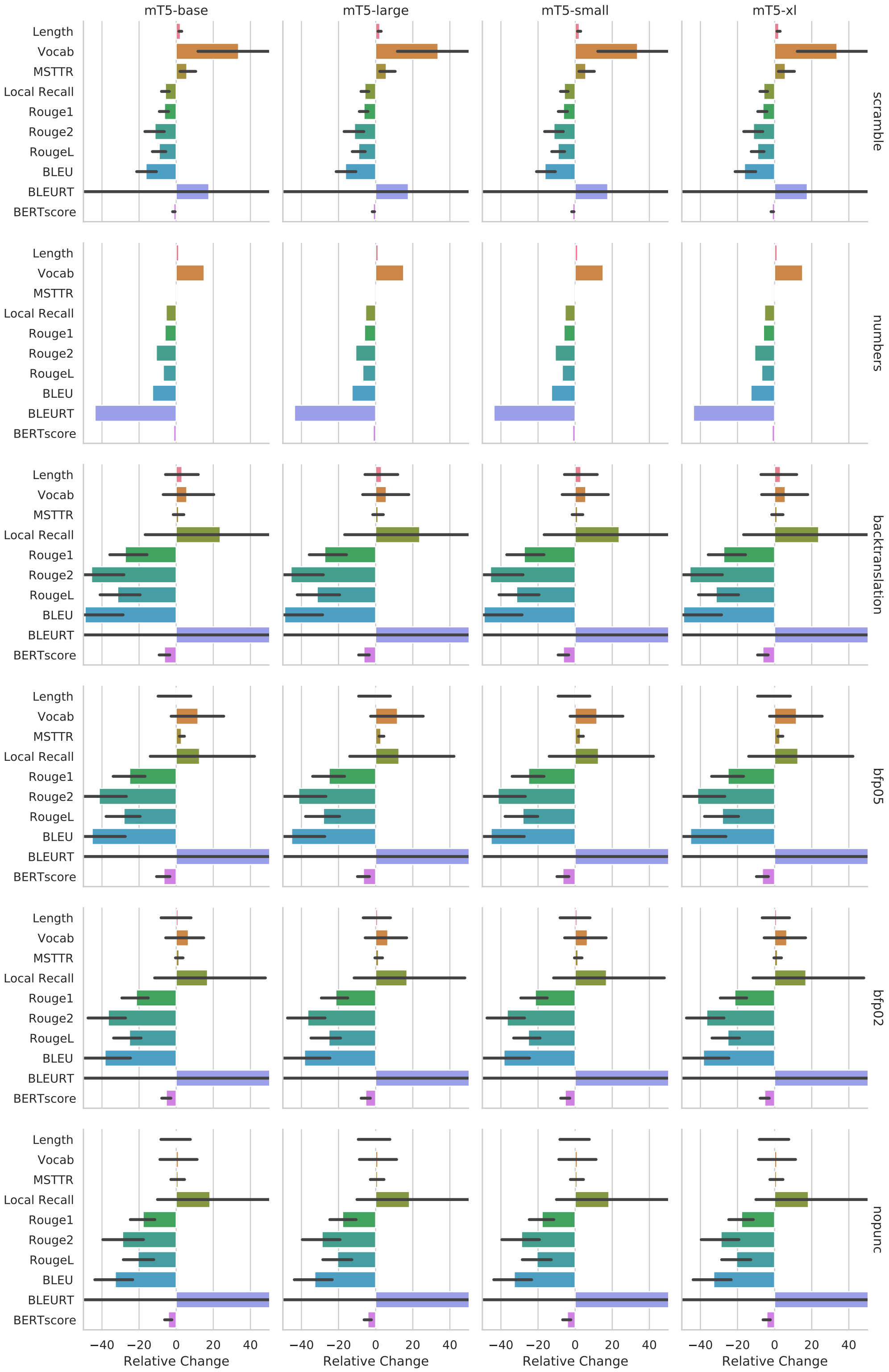

Figure 5: Extended version of Figure 2 of the main paper. On the x-axis, we present the relative metrics change across models as a result of applying transformations. All performance-related metrics tend to decrease while often, the diversity of output increases as a model relies on its LM more and focuses on the input less. Optimally, we would like to observe no difference in any of these cases. Note that the x-axes are cut off at -50% and 50% and some changes are beyond those thresholds.